Chapter 13

Market Models

In these notes we model the rates that are actually quoted rather than the instantaneous abstractions of the previous chapters, which makes the calibration instruments price exactly by construction and costs the low-dimensional state that made those chapters tractable. We build the Libor market model with its drift derived rather than quoted, do the same in swap rate coordinates, and then settle the choice between them — forwards and swap rates cannot both be lognormal, and how much that matters is a measurement rather than a preference. The answer reorders the modelling priorities: decorrelation is worth percentage points of volatility and the departure from lognormality is worth hundredths.

13.1 Modelling What Is Quoted

Chapter 8 gave the general framework for arbitrage-free curve dynamics and chapter 12 made it usable by forcing a low-dimensional Markov state. Both model the instantaneous forward rate $f(t,T)$, which is not a rate anybody trades. Every quoted instrument references an accrual over a real period — three months, six months — and the passage from the instantaneous object to the traded one is an approximation, small but present, in the pricing of every calibration instrument.

The market models invert that. They take as primitive exactly the rates the market quotes, and accept whatever dynamics follow. The reward is immediate: the calibration instruments price by the market's own formula, exactly, with nothing to approximate. The cost is the subject of most of this chapter.

Definition 13.1 (Tenor structure). Fix dates $T_0 < T_1 < \cdots < T_N$ with accruals $\delta_i = \tau(T_i, T_{i+1})$ on the appropriate convention of chapter 7. The forward rates are

$$F_i(t) = \frac{1}{\delta_i}\left(\frac{P(t,T_i)}{P(t,T_{i+1})} - 1\right), \qquad i = 0, \dots, N-1, \tag{13.1}$$

so F_i is the rate that fixes at T_i and pays at T_{i+1}. The curve and the vector $(F_0, \dots, F_{N-1})$ carry the same information: (13.1) inverts to

$$P(t,T_k) = P(t,T_{\eta(t)}) \prod_{j=\eta(t)}^{k-1} \frac{1}{1+\delta_j F_j(t)}, \tag{13.2}$$

with $\eta(t)$ the index of the next date in the structure.

One observation about (13.1) does all the structural work in this chapter.

Lemma 13.2 (Each forward is a martingale under its own measure)**.** *F_i is a martingale under the T_{i+1}-forward measure, the measure of chapter 6 whose numeraire is $P(\cdot, T_{i+1})$.*

Proof. Rearranging (13.1),

$$\delta_i F_i(t)\, P(t, T_{i+1}) = P(t, T_i) - P(t, T_{i+1}),$$

and the right hand side is a portfolio of two traded assets — long a bond maturing at T_i, short one maturing at T_{i+1}. So $\delta_i F_i$ is the value of that portfolio divided by $P(\cdot, T_{i+1})$, which is the numeraire, and chapter 6 showed that any traded asset divided by the numeraire is a martingale under the associated measure. □

Remark (Why that is more than a technicality)**.** Lemma 13.2 is not a property of a model. It holds in *any* arbitrage-free market, because it follows from (13.1) and nothing else. So a forward rate has a distinguished measure in which it has no drift, handed over by the definition, and the only remaining modelling choice is what its volatility is.

Choose that volatility to be deterministic and lognormal, $dF_i/F_i = \sigma_i(t)\, dW_i$, and a caplet on F_i — which pays $\delta_i(F_i(T_i) - K)^+$ at T_{i+1} — has value

$$P(0, T_{i+1})\, \delta_i\, \mathbb{E}^{T_{i+1}}\left[(F_i(T_i) - K)^+\right],$$

an expectation of a call payoff on a driftless lognormal variable, which is Black's formula exactly. Not approximately, and not after an adjustment. The market quotes caps in Black volatilities; this model's parameters *are* those volatilities.

That is the whole appeal: the model has been built so that its calibration is a change of notation. Every chapter before this one had to solve for parameters that reproduce quotes. Here the quotes are the parameters.

13.2 The Drift, and Why There Has To Be One

Lemma 13.2 gives each forward its own measure. A portfolio containing several of them has to be priced under *one*, and no measure makes them all driftless.

Theorem 13.3 (The terminal measure drift)**.** *Suppose $dF_i/F_i = \sigma_i\, dW_i$ under the T_{i+1}-forward measure for each i, with $d\langle W_i, W_j\rangle = \rho_{ij}\, dt$. Then under the terminal measure, with numeraire $P(\cdot, T_N)$,*

$$\frac{dF_i}{F_i} = -\sigma_i \sum_{j=i+1}^{N-1} \frac{\delta_j \rho_{ij} \sigma_j F_j}{1 + \delta_j F_j}\, dt + \sigma_i\, dW_i^N. \tag{13.3}$$

Proof. The change of measure from T_{i+2} to T_{i+1} is governed by the ratio of numeraires, and (13.1) says what that ratio is:

$$\frac{P(t, T_{i+1})}{P(t, T_{i+2})} = 1 + \delta_{i+1} F_{i+1}(t).$$

By the change of numeraire theorem of chapter 6, moving from numeraire A to numeraire B adds to a process's drift the covariation of that process with $\log(B/A)$ — for a positive process, the covariation of its logarithm with $\log(B/A)$. Here $\log(B/A) = -\log(1+\delta_{i+1}F_{i+1})$, whose stochastic differential has volatility

$$-\frac{\delta_{i+1}\,dF_{i+1}}{1+\delta_{i+1}F_{i+1}} = -\frac{\delta_{i+1}F_{i+1}\sigma_{i+1}}{1+\delta_{i+1}F_{i+1}}\,dW_{i+1},$$

so moving F_i from its own measure to the T_{i+2} measure adds the drift

$$-\sigma_i\frac{\delta_{i+1}\rho_{i,i+1}\sigma_{i+1}F_{i+1}}{1+\delta_{i+1}F_{i+1}}\,dt.$$

Iterating from T_{i+1} out to T_N accumulates one such term per step, which is (13.3). □

Remark (Reading the drift)**.** Three features of (13.3) explain most of what a market model is like to work with.

The sum is empty for $i = N-1$. The last forward is a martingale under the terminal measure, since its own measure *is* the terminal measure. This is the cheapest available check on any implementation, and an error in the limits of the sum shows up here before anywhere else.[1]

The drift is state-dependent. F_j appears on the right hand side, so (13.3) is not a lognormal process with a known drift; it is a coupled system of N equations in which every forward's drift depends on the current level of all the later ones. There is no closed form for the joint law — the drift is what enforces no-arbitrage between rates that were each specified separately, which is chapter 8's drift condition arriving in discrete coordinates.

The sign is one-directional. Every term is negative, so under the terminal measure every forward but the last drifts down, and further the further it sits from the numeraire. A numeraire at the far end of the structure gives the near forwards large drifts, which are exactly the forwards a short-dated product depends on.

Structure (The drift is a change of coordinates, not a force)**.** The algebra can make it look like a modelling assumption and it is not one.

Each forward was specified as driftless in its own measure. Nothing about the market has been asserted beyond that. The drift appears purely because the forwards are being written in a common coordinate system, and a change of measure is a change of coordinates on the space of processes: the same process described from a different numeraire acquires a drift, in the way that a straight line acquires curvature when written in polar coordinates.

13.3 The Numeraire Is a Choice, and Jamshidian's Is the Better One

The terminal measure is the easiest to derive and among the worst to simulate in. Two problems, and the first is fatal rather than inconvenient.

A payoff with cashflows beyond T_N cannot be priced at all, because there is no numeraire past the end of the structure. That sounds like an artefact of choosing N too small, and it is not: a

[1] `tested`, along with the drift growing with distance from the numeraire, and `a caplet repricing to Black` — which only works if the drift and the deflator cancel, and they cancel only if both are right.

callable structure whose exercise generates a swap extending past the last modelled date requires the bond $P(t,T)$ for $T > T_N$, which (13.2) cannot supply. Extending N moves the problem rather than solving it, because the terminal forwards then have no volatility data to calibrate against.

The second is that the drifts of the near forwards are the largest, by the sign argument above, so the discretisation error is concentrated exactly where a short-dated product lives.

Definition 13.4 (The spot Libor measure)**.** Let B be the value of a unit invested at T_0 and rolled at every date in the structure at the rate fixing then:

$$B(T_k) = \prod_{j=0}^{k-1} \bigl(1 + \delta_j F_j(T_j)\bigr), \tag{13.4}$$

interpolated between dates by the bond $P(t, T_{\eta(t)})$. Jamshidian's *spot Libor measure* is the measure associated with B as numeraire.

This is a discretely rebalanced money market account, and it is the natural numeraire for a market model in the same way that the continuously compounded account is natural for a short rate model — with the difference that (13.4) is built entirely from rates the model already has, so it requires nothing that is not modelled.

Remark (What changes)**.** Repeating the change of numeraire argument with B in place of $P(\cdot, T_N)$ gives

$$\frac{dF_i}{F_i} = \sigma_i \sum_{j=\eta(t)}^{i} \frac{\delta_j \rho_{ij} \sigma_j F_j}{1 + \delta_j F_j}\, dt + \sigma_i\, dW_i^B, \tag{13.5}$$

and the differences from (13.3) are exactly the ones wanted. The sum now runs *forward* from the front of the structure to i, so it involves only forwards that have not yet fixed — and it is empty for the front forward, which is therefore the one with no drift, rather than the far one. The sign is positive, the numeraire never expires, and a product with cashflows at many dates is priced by discounting each on the same path of B rather than by deflating everything to a single far date.

The cost is that B is path-dependent: (13.4) depends on the rates that fixed along the way, so it cannot be read off the current state. For a simulation that is no cost at all, since the path is being generated anyway. For anything wanting a Markov state it would be, which is another way of seeing that this model has given up on having one.

13.4 No State Variable

That giving-up is the central cost of the whole construction.

Chapter 12 went to considerable trouble to force the curve's evolution into a low-dimensional Markov state, because a low-dimensional state is what permits a partial differential equation, and a partial differential equation is what permits early exercise to be handled by backward induction. The market model has no such state. Its state is the whole vector $(F_{\eta(t)}, \ldots, F_{N-1})$ — forty numbers for a ten year quarterly structure, eighty for twenty years — and chapter 19 measures what a grid does in forty dimensions.

So a market model is a simulation model, necessarily and not by preference. Everything follows from that:

- European payoffs are straightforward: simulate to the expiry, average the payoff over paths.
- Early exercise requires the continuation value at each exercise date as a function of a forty-dimensional state, which must be estimated by regression on the simulated paths, and the choice of regressors is a modelling decision that no calibration constrains. Chapter 19 is where that difficulty is treated; here the point is that the model's structure forces it.
- Greeks come from differentiating a simulation, with the accuracy problems chapter 19 sets out, rather than from reading a derivative off a grid.

Structure (The trade that this chapter is)**.** Set the two families side by side and the trade is clean.

A short rate or quasi-Gaussian model has a small Markov state, so it admits a partial differential equation, backward induction, and exact early exercise — and it prices its calibration instruments approximately, with a volatility structure constrained by the requirement that the state stay small. Chapter 12 measures what that constraint costs.

A market model prices its calibration instruments exactly and imposes no constraint on the volatility structure — and it has no state, so early exercise becomes a regression problem whose error is not controlled by anything the calibration sees.

Neither is better. The choice is which error you would rather have, and that is decided by the product: the exact-calibration model is right when the answer is dominated by the vanilla prices, and the small-state model is right when it is dominated by the exercise decision. Chapter 25 returns to this, because it is the clearest instance in the book of a compromise that was forced by compute and is now a genuine choice.

13.5 The Same Curve in Swap Rate Coordinates

Nothing in the construction was specific to forward rates. The argument was: find a quantity that is a traded asset divided by a numeraire, declare it lognormal in that numeraire, and the corresponding option prices by Black. Swap rates satisfy the same description.

Lemma 13.5 (A swap rate is a martingale under its annuity measure)**.** *The par swap rate $S_{a,b}$ of chapter 7 is a martingale under the measure whose numeraire is the annuity $Ann_{a,b}(t) = \sum_{k=a+1}^{b} \delta_{k-1} P(t, T_k)$.*

Proof. The par rate is defined by the fixed leg matching the floating leg, which gives

$$S_{a,b}(t)\,\mathrm{Ann}_{a,b}(t) = P(t, T_a) - P(t, T_b),$$

again a portfolio of two bonds. So $S_{a,b}$ is a traded asset divided by the annuity, and the annuity is a positive portfolio of bonds and hence an admissible numeraire. □

Declaring $dS_{a,b}/S_{a,b} = \sigma_{a,b}\,dW$ under that measure makes a swaption — which pays $\mathrm{Ann}_{a,b}(T_a)(S_{a,b}(T_a)-K)^+$ — price by Black exactly, with the annuity as the discount factor. That is the swap market model, and for the *co-terminal* family $S_{a,N}$, $a = 0, \ldots, N-1$, it plays the same role for swaptions that the Libor market model plays for caps.

Theorem 13.6 (Both cannot be lognormal)**.** *If every forward F_i is lognormal under its own measure, then the swap rates are not; and if the co-terminal swap rates are lognormal, the forwards are not.*

Proof. By (13.2) and Lemma 13.5,

$$S_{a,b} = \frac{P(t,T_a) - P(t,T_b)}{\sum_k \delta_{k-1} P(t,T_k)} = \sum_{i=a}^{b-1} w_i(F)\, F_i, \tag{13.6}$$

where the weights w_i are ratios of bonds and therefore themselves functions of the forwards. So a swap rate is a weighted sum of the forwards, and a weighted sum of lognormal variables is not lognormal — the family is not closed under addition, as it is under multiplication. The same argument run backwards inverts the conclusion. □

So the two models are not two descriptions of one thing. They are two incompatible specifications, and at most one of them can price both caps and swaptions by the market's formula. The choice of coordinates is a choice of which set of quotes to get exactly right.

13.6 Which Coordinates, and What the Answer Costs

Theorem 13.6 says there is an approximation. It says nothing about its size, and the size is what decides the question, so it has to be measured.

Calculation 13.7 (What a forward-based model costs on swaptions)**.** Take a flat curve at 4% on ten semiannual periods, every forward lognormal at 25% volatility, correlations decaying exponentially with the gap between maturities,

$$\rho_{ij} = e^{-\beta_\rho |T_i - T_j|}, \tag{13.7}$$

so β_ρ is a decay rate per year: at $\beta_\rho = 0$ every forward moves with every other, and a large β_ρ leaves forwards a few years apart nearly independent. The forwards have no smile whatever, by construction. Simulate, price swaptions on the five year swap starting in two years across strikes from 70% to 130% of the forward, and invert to Black implied volatility.[2]

The induced smile spans under a tenth of a volatility point. At a hundred thousand paths the shape is not stable from one seed to the next. Running two million paths over several seeds a consistent shape does appear, implied volatility rising with strike by about three hundredths of a point across that strike range. Resolving it took eight million paths.

The relevant comparison is a swaption bid-offer, which is a large fraction of a volatility point. So the approximation is one to two orders of magnitude below the price at which the instrument can be traded.

Calculation 13.8 (And what the correlation is worth)**.** The same model, at the money, with the volatility of the swap rate read off its own option price rather than assumed.[3]

Every forward carries 25% volatility. The swap rate averaging them comes out near 22.8%, because (13.6) is an average of imperfectly correlated things and an average moves less than its parts. Increasing the correlation decay — decorrelating the forwards further — lowers it again, measurably.

So the correlation structure moves the at-the-money swaption volatility by whole percentage points, while the departure from lognormality moves the smile by hundredths of one.

[2] `measured`.
[3] `measured`.

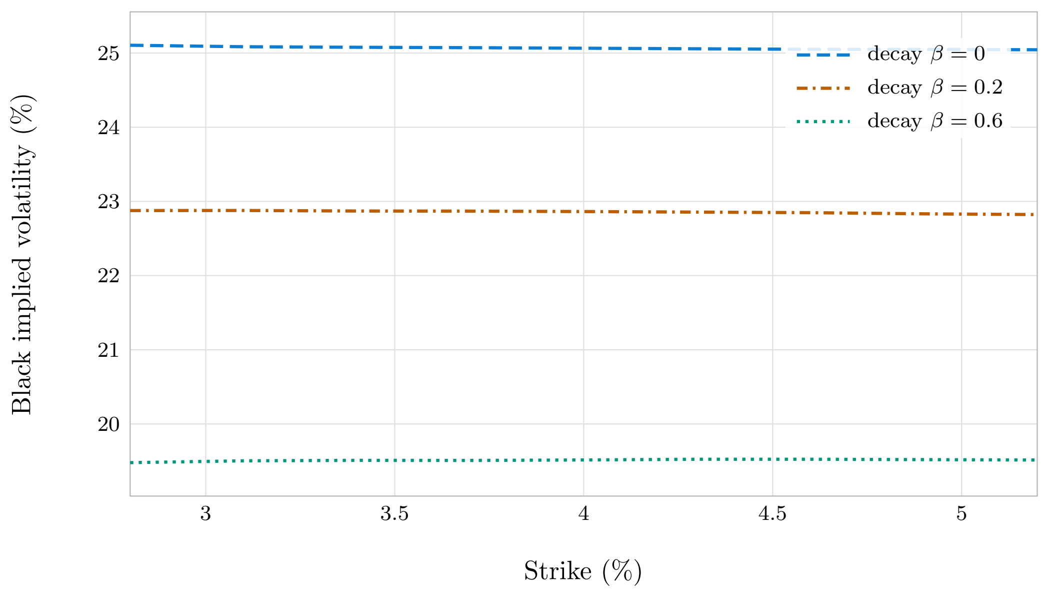

Figure 13.1: The swaption smile at three correlation decays, with every forward carrying the same 25% volatility so that nothing on the picture comes from the inputs. At $\beta_\rho = 0$ the forwards move together and the swap rate inherits their volatility exactly, which is the internal check: 25.1% against a 25% input. Decorrelating them takes the level to 22.9% and then 19.6%, while each curve stays flat across strikes to within a twentieth of a point. The vertical spread between the curves is what the correlation is worth; the tilt within each is what the departure from lognormality is worth, and the ratio of the two is why this chapter argues the correlation first.

Drawn by `Lmm::simulate` and `swaption_smile`.

Remark (Two things these numbers do not say)**.** *They do not say the model produces a smile.* A tenth of a volatility point is not a smile mechanism; it is the residue of a coordinate mismatch, measured in a model built to have no smile at all — every forward is lognormal by construction. A lognormal market model cannot fit a market smile in either set of coordinates, which is what §13.9 is about. What the calculation says is narrower and more useful: the *approximation* incurred by choosing one coordinate system over the other is not the thing standing between the model and the market. The smile has to be put in separately, and doing so is orthogonal to the question this section is asking.

They do not say the correlation is unobservable. It is quoted, if indirectly. Because the swap rate's variance is the double sum (13.9), every swaption price is a statement about ρ, and the next section calibrates it that way: caps pin the diagonal and swaptions pin the off-diagonal. What makes the correlation the largest risk in the model is not that the market is silent about it but that the market does not say *enough* — a hundred or so liquid swaptions against the $N(N-1)/2$ entries of a full matrix — so the fit proceeds under a low-rank restriction that is an assumption rather than a measurement.

Structure (The measurement reorders the priorities)**.** Put Calculation 13.7 and Calculation 13.8 together and they answer a question that is usually argued about on grounds of taste.

The thing a swap market model buys — exact pricing of its own calibration instruments — is worth hundredths of a volatility point. The thing both models must get right, and which neither gets for free, is the correlation between rates, and that is worth percentage points. The two differ by two orders of magnitude, so a decision made on the first while ignoring the second is a decision made on the wrong axis.

That does not make the choice of coordinates irrelevant, but it relocates the argument. It is not about calibration accuracy. It is about which coordinates make the *payoff* simple:

- A Bermudan swaption's exercise decision at each date compares a swap rate with a strike. In co-terminal swap rate coordinates that is a comparison of one state variable with a number; in forward coordinates it is a comparison of a nonlinear function of many state variables with a number, and the regression that estimates the continuation value has to learn that function. So the swap market model makes the exercise boundary simple, which is where the uncontrolled error of the previous section lives.

- A product touching several tenors, or a book that must be priced consistently across caps, swaptions and exotics, cannot use co-terminal swap rates at all: the family is tied to one terminal date, and a second product with a different terminal date needs a different, mutually inconsistent, model. The forward family spans everything by (13.2).

Which gives the criterion. Use swap rate coordinates for a book of Bermudan swaptions on one tenor structure, where the exercise decision is the whole difficulty and the calibration is a read-off. Use forward coordinates for anything that has to span tenors, and pay a cost that Calculation 13.7 bounds at a fraction of a bid-offer. And in either case spend the modelling effort on the correlation, because Calculation 13.8 is where the money is.

13.7 Calibration

What has to be chosen is the volatility function $\sigma_i(t)$ for each forward and the correlation ρ_{ij} between them. The instruments available are caps, which price each forward on its own, and swaptions, which price combinations.

A caplet on F_i depends on σ_i only through

$$\int_0^{T_i} \sigma_i(t)^2 \, dt, \tag{13.8}$$

the accumulated variance to its own expiry, and on no correlation at all. A swaption depends on the whole matrix, since by (13.6) the swap rate's variance is

$$\sigma_{a,b}^2 \approx \frac{1}{S_{a,b}^2} \sum_{i,j} w_i w_j F_i F_j \rho_{ij} \int \sigma_i \sigma_j \, dt. \tag{13.9}$$

So caps pin the diagonal and swaptions pin the off-diagonal. That is a clean identification statement of the kind chapter 20 asks for, and it is the reason both instrument sets are needed: caps alone leave the correlation entirely free, which by Calculation 13.8 is the parameter that matters most.

Remark (Time-homogeneity is not a convenience)**.** Equation (13.8) constrains only the integral, so infinitely many $\sigma_i(\cdot)$ fit the same caplet. The standard resolution is to insist that the volatility depend on time to maturity rather than on calendar time,

$$\sigma_i(t) = \phi_i \, g(T_i - t), \tag{13.10}$$

with a shared shape g and a per-forward scaling ϕ_i absorbing the fit.

This is usually presented as a parsimonious parametrisation. It is more than that, and chapter 10 has already made the argument in another setting. A calendar-time volatility fitted to today's caps will generally have σ_i collapsing as t approaches T_i in whatever way the fit requires, and that collapse is a statement about the volatility structure the model will have *tomorrow*. A model whose implied volatility surface a year from now is nothing like today's cannot be used for anything whose value depends on the surface at a future date — which is precisely what a callable is. Time-homogeneity is the discipline that stops the calibration from spending the future to fit the present, and it is the same discipline that chapter 10 demanded of a stochastic volatility model.

Remark (The correlation matrix has too many entries)**.** A full ρ has $N(N-1)/2$ parameters, which for a forty-forward structure is seven hundred and eighty, against perhaps a hundred liquid swaption quotes. The market does not identify it, and chapter 20's warning applies exactly: a fit that succeeds does not mean the parameters were determined.

What is done is to impose a low-rank form, $\rho = LL^\top$ with L of d columns. This is the truncated spectral decomposition of chapter 12, where it is set out in full: writing $\rho = \sum_j \lambda_j v_j v_j^\top$ and keeping the largest d terms gives L with columns $\sqrt{\lambda_j}\, v_j$, and no matrix of rank d approximates ρ better. It is the same reduction that chapter made for a different reason and which chapter 24's factor analysis of real curves justifies — three or four factors carry almost all the variance. The parametrisation used above, $\rho_{ij} = e^{-\beta|T_i - T_j|}$, is the one-parameter version: it has the right qualitative shape, decaying with separation, and one number to fit.

What is the same as is chapter 12 the algebra: a symmetric positive semi-definite matrix, its spectral decomposition, and a truncation that is optimal at its rank. What differs is three things. That chapter decomposes a *covariance* estimated from a history of curve moves, so its eigenvectors are a measurement and the truncation discards information that was in the data. Here the matrix is a *correlation*, because the volatilities are carried separately and calibrated from caps, and it is not estimated from anything — it is a parametric shape, or the output of a fit to swaptions. So the truncation is not discarding a measurement; it is the reason the fit is identified at all, since the full matrix has more entries than there are prices.

And a correlation matrix does not survive truncation as cleanly as a covariance does. Keeping d terms of $\sum_j \lambda_j v_j v_j^\top$ leaves a matrix whose diagonal entries are below one, since each is $\sum_{j\le d} \lambda_j v_{j,i}^2$ and the discarded terms were positive. A correlation matrix with a diagonal below one is not a correlation matrix, so the rows of L are rescaled to unit length afterwards. That restores the diagonal and perturbs every off-diagonal entry slightly, and the perturbation — not the truncation — is what a low-rank correlation actually costs.

One thing does carry across unchanged. A correlation between diffusions is a property of the diffusion coefficients, which chapter 6 showed a change of measure leaves alone, so a shape estimated from history is an admissible input to a pricing model in a way that a drift estimated from history would not be. That is why desks are willing to take the *shape* of the correlation from a historical decomposition and fit only its overall level to swaption prices — and it is the same argument that licensed measuring a backbone from a time series in chapter 10.

The choice of d then has a consequence that is easy to miss. A d-factor model can produce at most d independent shapes of curve move, so a payoff sensitive to a $(d+1)$th shape is priced as though that shape did not exist. This is chapter 12's measurement in market model coordinates, and the resolution is the same: choose d by what the payoff is exposed to, not by what fits.

13.8 Simulating It

Three practical points, each of which is a place the implementation can be wrong while appearing to work.

Use logarithms. Simulating $\log F_i$ keeps each forward positive by construction and makes the diffusion term exact, leaving discretisation error only in the drift.

The drift must be frozen, and that is the scheme's only error. Equation (13.5) depends on the current forwards, so a step requires the drift at some point within it, and using its value at the start is the standard choice. The resulting bias is not detectable by looking at the output; it is detectable by the caplet test, which compares a simulated price against a closed form that shares none of the simulation's machinery. That test is the reason to trust the implementation, and no amount of internal consistency substitutes for it.

Check the martingales. Every forward has a measure in which it is driftless, and under the simulated measure the front forward in (13.5) and the last in (13.3) are driftless outright. A simulation whose martingales drift has a bug, and the check costs one line.

Remark (Where the real error is)**.** None of the above is the dominant error in a Bermudan price. That is the regression for the continuation value.

The continuation value is a function on a forty-dimensional state space, estimated from paths. Whatever basis is chosen, it is a low-dimensional approximation of a high-dimensional function, and its error is a bias in the exercise decision — exercising when continuing was better, or the reverse. The bias is one-signed in a way that flatters the price, because a suboptimal exercise rule always produces a lower value, so the simulated price of a Bermudan is biased *low* by an amount nobody can compute directly.

Which returns to the trade of this chapter. The small-state models of chapter 12 handle the exercise decision exactly and the calibration approximately. This model handles the calibration exactly and the exercise decision approximately, and the approximation in the exercise decision is the larger of the two for exactly the products this model exists to price.

13.9 What Is Missing, and It Is the Smile

Everything above had deterministic volatility. So every caplet in the model has one Black volatility per expiry, and the market quotes a different volatility at every strike. The model as constructed cannot fit the market at all across strikes: it does not fit the smile badly, it has no smile.

The historical repair was to make the volatility depend on the level of the forward — a displaced diffusion, $d(F_i + \alpha)/(F_i + \alpha) = \sigma_i \, dW_i$, or a CEV exponent. This produces a skew and it is exactly what chapter 9 calls a local volatility model, applied to each forward. So it inherits, precisely, the pathology chapter 10 measures: the skew is glued to the level of rates, the backbone is deterministic, and forward smiles flatten as expiry recedes.

Why not mark each swap rate as SABR?

There is an obvious construction to try, and it repays working through because it is both used and more limited than it looks.

The swaption cube is marked in SABR: for every expiry and tenor, a desk carries four parameters that reproduce that smile. Lemma 13.5 says each co-terminal swap rate is a martingale under its own annuity measure. So declare each one to follow SABR dynamics there. Every swaption then prices by the market's own formula with the market's own parameters, and the calibration is not a fit at all — it is a copy of the marks.

Three things are wrong with concluding that the model is therefore right.

A perfect fit instrument by instrument is not a joint model.

The issue is counting, and it bites on the cube rather than on the co-terminal family. The $n-1$ co-terminal rates are a coordinate system for a curve with $n-1$ degrees of freedom — the overall level divides out of every one of them — so giving each its own diffusion is exactly determined and not over-determined. A desk does not mark $n-1$ rates. It marks a cube, every expiry against every tenor, which is many more rates than the curve has dimensions; and since all of them are smooth functions of the same bonds, Itô expresses the diffusion of each in terms of the others. Most of the cells are therefore consequences rather than choices, and marking every one of them independently is Theorem 13.6 again with more instruments: the marks can be mutually inconsistent, and generically are.

The parameters are constants, so the smile moves wrongly. This is the more serious objection and it is chapter 10's, transplanted. SABR's β — the backbone exponent, not the correlation decay β_ρ of (13.7) — fixes the backbone deterministically, and ρ and ν do not move at all, so the model's smile a year from now is determined by today's marks in a way the market's smile does not respect. For a European swaption that is a second-order complaint, because the price depends on the terminal distribution and the terminal distribution was fitted. For a callable it is first order: the exercise decision is priced off the smile *at the exercise date*, so fitting every smile today and moving them incorrectly is being wrong about the only quantity the construction exists to compute. Fitting the cube exactly buys the vanillas, and the vanillas were already quoted.

And it leaves free the parameter that matters most. A SABR specification per swap rate says nothing whatever about the correlation between different swap rates, or between their volatilities. Those are extra assumptions, unconstrained by any of the marks that were copied — and Calculation 13.8 measured that the correlation is worth percentage points of volatility while the smile fit is worth hundredths. So the construction spends all its structure on the cheap axis and leaves the expensive one to be assumed.

Remark (One further reason the fit is not as exact as it appears)**.** Hagan's SABR formula is an asymptotic expansion, not a solution, and it is known to admit arbitrage at low strikes — the implied density it corresponds to can go negative. So a market marked in SABR is marked with a convention that is not itself arbitrage-free, and a model that reproduces those marks exactly reproduces the arbitrage along with them.

The honest verdict is that this is a good *marking* device and not a model of a callable. Carrying the cube as a field of SABR parameters is an excellent way to interpolate a surface, reprice vanillas quickly, and communicate a smile in four numbers. What it does not supply is either of the two things a Bermudan needs — how the smile evolves, and how the rates move together — and those are precisely the two it leaves unspecified.

Which points at what the fix looks like. Rather than a separate SABR per rate, attach a *common* stochastic volatility process to the whole family, scaling the loadings of every rate together. Then the smile dynamics are specified once and jointly, the correlation structure is explicit rather than implied by a collection of marginals, and the model has as many smile parameters as the market can identify instead of four per cell of a cube. That is the stochastic-volatility market model, and it is what the remark below is about.

The stochastic-volatility market model

Definition 13.9 (Market model with a common variance factor)**.**

$$\begin{aligned} \frac{dF_i(t)}{F_i(t)} &= \mu_i(t)\,dt \;+\; \sqrt{V_t}\,\sigma_i(t)\,dW_t^i, \\ dV_t &= \theta\,(1 - V_t)\,dt \;+\; \eta\sqrt{V_t}\,dZ_t, \qquad V_0 = 1, \end{aligned} \tag{13.11}$$

with $dW^i dW^j = \rho_{ij}\,dt$, and μ_i the no-arbitrage drift of Theorem 13.3 in whichever measure is being simulated.

Four things about (13.11) are the whole of its design.

One factor, not one per rate. A single V multiplies every forward's volatility, so the smile dynamics are one specification rather than a field of them. That is the repair to the first objection above: the swap rates can no longer be given mutually inconsistent smiles, because they are not given smiles at all — they inherit one.

The variance is dimensionless. It starts and reverts to one, so $\sigma_i(t)$ keeps its old meaning as the level of volatility and η adds only randomness around it. Setting $\eta = 0$ returns the deterministic model exactly.[4]

The drift is not made harder. The drift is a sum of products of *two* volatilities — it comes from a covariation — so it carries V where each diffusion carries $\sqrt{V}$, and the derivation of Theorem 13.3 goes through with V_t multiplying the same sum. Nothing has to be rederived; a factor is threaded through.

And the split of labour is clean. $\sigma_i(t)$ fits the caplet levels as before, η and θ fit the caplet smile's curvature and how it flattens with expiry, and ρ_{ij} still fits the swaption grid. The instruments that identified the deterministic model identify this one, with the smile parameters answering to the strike dimension nobody was using.

Calculation 13.10 (What the factor is worth, against what the coordinates were worth)**.** The same ten-period structure as Calculation 13.7, every forward at 25%, correlation decay $\beta_\rho = 0.2$, priced across strikes from 70% to 130% of the forward swap rate.

At $\eta = 0$ the smile spans four hundredths of a volatility point, which is Calculation 13.7 again. At $\eta = 0.6$ it spans four tenths, and at $\eta = 1.2$ about one and a half points — stable across seeds and path counts.[5]

So one extra state variable moves the smile by more than an order of magnitude beyond anything the choice of coordinates could reach, and into the range where swaptions actually trade. The at-the-money level drifts down slightly as η grows, from 22.9% to 21.6%, which is the usual convexity of averaging over a random variance and is absorbed by recalibrating σ_i.

A single common factor is a strong assumption.

Every smile moves together. One V cannot decorrelate the two year's smile from the ten year's, and a product sensitive to that — a spread option on volatility, or a callable whose exercise depends on how the front and back of the surface move apart — is priced as though the question did not arise. This is chapter 12's "d factors give d shapes" arriving in the volatility dimension rather than the rate one, and the repair is the same: more factors, at the usual cost.

And (13.11) as written has a symmetric smile. The skew has to come from somewhere else — a displaced diffusion or a CEV exponent on each forward, or a correlation between Z and the W^i — and which of those is chosen is exactly chapter 10's backbone question, asked again in a market model. The measurement of chapter 10 applies unchanged: the exponent is observable, and it is not any of the conventional choices.

Remark (Why that matters more here than for a vanilla)**.** Chapter 10 showed that a local volatility model fits today's smile exactly and gets the future smile wrong. For a European option that is a second-order complaint: the price depends on the terminal distribution, which is fitted.

For a callable it is a first-order complaint. The holder's decision at a future exercise date compares the swap against the value of continuing, and the value of continuing is the value of the remaining option — which depends on the volatility surface *at that date.* The forward smile is not a diagnostic here; it is an input to the price. A model that gets it wrong prices the exercise decision wrong, which is the one quantity the whole apparatus of this chapter was built to get right.

[4] `checked to machine precision`.

[5] `measured`.

So the answer to whether a market model needs stochastic volatility is yes, and the argument is not the usual one about fitting the smile better. It is that a deterministic-volatility model, whether it has a skew or not, has a wrong forward smile, and a callable is priced off the forward smile. The stochastic volatility versions — a variance process multiplying the loadings, in the manner of chapter 11 — are what fix it.

With one ordering caveat, which Calculation 13.8 supplies. The smile is second order in size to the correlation structure, so a model with rich volatility dynamics and a one-factor correlation is worse than the reverse. Get the number of factors and their correlation right first; then the smile. Both, now that compute forces neither to be sacrificed — which is chapter 19's argument arriving at a specific recommendation.

References

- Brace, A., Gatarek, D., & Musiela, M. (1997). The market model of interest rate dynamics. *Mathematical Finance*, 7(2), 127–155.

- Jamshidian, F. (1997). Libor and swap market models and measures. *Finance and Stochastics*, 1(4), 293–330.

- Miltersen, K. R., Sandmann, K., & Sondermann, D. (1997). Closed form solutions for term structure derivatives with log-normal interest rates. *Journal of Finance*, 52(1), 409–430.

- Rebonato, R. (2002). *Modern Pricing of Interest-Rate Derivatives: The LIBOR Market Model and Beyond*. Princeton University Press.

- Andersen, L., & Andreasen, J. (2000). Volatility skews and extensions of the Libor market model. *Applied Mathematical Finance*, 7(1), 1–32.

- Piterbarg, V. (2005). Stochastic volatility model with time-dependent skew. *Applied Mathematical Finance*, 12(2), 147–185.

- Andersen, L., & Piterbarg, V. (2010). *Interest Rate Modeling*. Atlantic Financial Press.